Historiallinen paikkatieto semanttisessa webissä: Biografiasampo.fi

Eero Hyvönen

Semanttinen web tarjoa lupaavan lähestymistavan tiedon esittämiseen historiallisissa paikannimirekistereissä ja rekisterien julkaisemiseen verkkopalveluina sekä ihmisten että koneiden käyttöön. Esimerkki älykkäästä digitaalisesta kulttuurihistoriallisesta sovelluksesta on Suomen kansallisia elämäkertakokoelmia julkaiseva ja rikastava Biografiasampo.fi.

Historiallisen paikkatiedon käsittely on haasteellista: ajan kuluessa paikkojen nimet vaihtuvat, paikkojen sijainti, alue ja tyyppi voivat muuttua ja paikka voi kokonaan hävitä muinaisen Karthagon tapaan. Lisäksi tietomme historiallisista paikosta on epävarmaa (käytiinkö taistelu Vuoksen tuolla tai tällä puolella?), luonteeltaan sumeaa (mistä kohdasta alkaa Korvatunturi?) ja puutteellista (missä on antiikin maantieteilijän Pytheaan tarkoittama Ultima Thule?). Tilannetta ei helpota se, että paikoista on yleensä ollut käytössä monenlaisia ja erikielisiä nimiä.

Historiallisten paikkojen ontologiat ja ontologiapalvelut

Semanttisessa webissä[1] paikannimirekisterit (gazetteer) julkaistaan ns. ontologioina. Niissä jokaiselle paikalle on luotu yksilöivä HTTP URI -tunniste, johon liitetyt ominaisuudet määrittelevät paikan tiedot kuten eri nimet, tyypin, koordinaatit, laajemman alueen ja paikan osat. Paikat ominaisuuksineen muodostavat verkkomaisen hierarkkisen tietämysgraafin (knowledge graph), joka esitetään semanttisen webin perustana olevan RDF-tietomallin ja standardien avulla[2].

Olennainen idea on, että HTTP URI -tunniste toimii sekä yksilöivänä tunnisteena että URL-osoitteena, josta käyttäjä voi saada lisätietoa ko. paikasta Tim Berners-Leen linkitetyn datan periaatteiden mukaisesti. Se, mitä tietoa paikkaontologiapalvelusta milloinkin palautetaan, esimerkiksi paikan kuvaava RDF-muotoinen data selaimen JavaScript-sovellukselle tai tästä ihmisen tutkittavaksi muodostettu HTML-sivu, hoidetaan HTTP-protokollan mukaisen sisältöneuvottelun (content negotiation) avulla.

Data rikastuu linkittämällä ja päättelemällä

Ontologiapalveluiden keskeisenä tavoitteena on edistää paikkatietoaineistojen yhteentoimivuutta. Eri datalähteistä saatava paikkatieto saadaan linkittymään toisiinsa käyttämällä ontologioissa yhteisesti sovittuja tai niihin sillattuja (map, align) tunnisteita.

Esimerkiksi Wikidatan[3] ja Wikipedioista louhitun DBpedian[4] linkitetyn datan palvelut sisältävät valtavan määrän paikkoja, jotka on sillattu muun muassa eri maiden paikannimirekisteristä tuotettuun GeoNames-rekisteriin[5] ja eri sovellusalojen paikkatietoaineistoihin maailmanlaajuisessa Linked Open Data -pilvessä[6]. Ontologioiden toinen tavoite on datan sisällön rikastaminen datajoukkojen keskinäisellä linkityksellä sekä päättelyllä, joka perustuu semanttisen webin standardien taustalla olevaan formaaliin predikaattilogiikkaan.

---

[1] E. Hyvönen: Semanttinen web. Linkitetyn avoimen datan käsikirja. Gaudeamus, 2018.
[2] https://www.w3.org/standards/semanticweb/
[3] https://www.wikidata.org/
[4] http://dbpedia.org
[5] https://www.geonames.org/
[6] http://linkeddata.org/

Kansallisbiografia Biografiasammon ytimessä

Syksyllä 2018 julkistetussa semanttisen webin Biografiasampo.fi-palvelussa[7] ydinaineistona on Suomalaisen Kirjallisuuden Seuran (SKS) toimittama Kansallisbiografia ja muut 13 100 elämäkertaa, joista on louhittu tekoälyn kieliteknologian ja tietämyksen irrotusmenetelmien (knowledge extraction) avulla 120 miljoonaa yhteyttä sisältävä linkitetyn datan tietämysgraafi[8].

Paikkaontologian perustaksi otettiin Kansalliskirjaston YSA-sanastosta irrotetuista paikannimistä muodostettu ontologia YSO-paikat[9], jonka fragmentaarisia hierarkioita muokattiin, täydennettiin ja laajennettiin mm. Sotasampo.fi-palvelun luovutetun Karjalan paikoilla. Paikkaontologia yhdistää monenlaisia Biografiasammossa yhdistyviä, toisiaan rikastuttavia aineistoja. Hierarkkista paikkaontologiaa on hyödynnetty mm. älykkäässä fasettihaussa, jossa vaikkapa ”Saksassa” syntyneitä henkilöitä etsittäessä löytyy Friedenaussa syntynyt sodanaikaisen Päämajan saksalaisen yhdysesikunnan kenraali Waldemar Erfurth, koska Friedenau on osa Berliiniä, joka puolestaan on osa Saksaa.

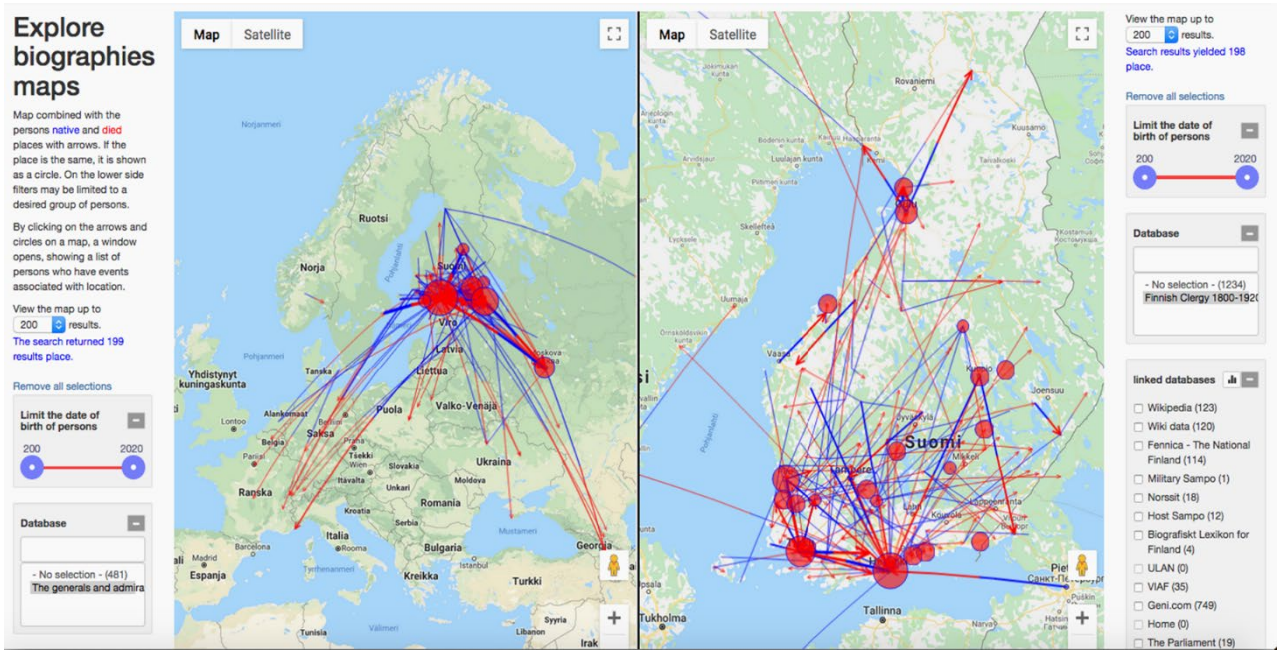
Järjestelmään sisältyvässä Yhteyshaku-sovelluksessa tavoitteena on päätellä ja löytää henkilöiden ja paikkojen välisiä yhteyksiä ja muodostaa niille suomenkieliset selitykset. Esimerkiksi fasettivalinnoilla ”Italia” ja ”taidemaalari” löytyy tieto, että Elin Danielson-Gambogi vastaanotti Firenzen kaupungin taidepalkinnon vuonna 1899 ja että ”Robert Wilhelm Ekman on luonut vuonna 1844 taideteoksen 'Maisema Subiacosta', joka kuvaa paikkaa Italia”. Jälkimmäisessä tapauksessa yhteys on muodostunut Biografiasampoon yhdistettyjen Ateneumin taidemuseon kokoelmatietojen kautta. Ateneumin lisäksi yhteyshaussa hyödynnetään kokoelmadataa kansallisbibliografia Fennicasta[10], Kirjasammosta[11], historiallisia tapahtumia kuvaavasta HISTO-ontologiasta[12] sekä J. V. Snellmanin kootuista teoksista[13], joista luotiin linkitetyn datan versio.

Visualisointeja paikkaontologialla

Paikkaontologian avulla tietoa voidaan myös visualisoida kartoilla. Mukaan on liitetty historialliset Karjalan kartat ja Senaatin kartasto. Esimerkiksi suomalaisia sotilaita ja pappeja kuvaavat ryhmät kartalla on muodostettu kahdella rinnakkaisella fasettihaulla Biografiasammon elämäkarttojen vertailunäkymässä, jossa elämä kuvataan sinipunaisena nuolena syntymäpaikasta kuolinpaikkaan.

Yhdellä vilkaisulla selviää, että sotilaat liikkuivat pappeja kansainvälisemmin ja vanhemmiten kohti etelää kuten eläkeläiset nykyään. Yhtä kaarta kartalla klikkaamalla pääsee käsiksi kaareen liittyviin elämäkertoihin. Esimerkiksi poikkeava elämänlanka Oulusta Länsi-Siperiaan osoittautuu Siperiaan maanmittaustöiden johtajaksi nimitetyn Gustav Adolf Silverhjelmin aiheuttamaksi.

[7] http://biografiasampo.fi/ ja
[8] https://seco.cs.aalto.fi/projects/biografiasampo/
[9] https://finto.fi/yso-paikat/fi/
[10] https://www.kansalliskirjasto.fi/fi/palvelut/metadatan-muunto-ja-valityspalvelut/avoin-data
[11] https://www.kirjasampo.fi/
[12] https://seco.cs.aalto.fi/ontologies/histo/
[13] http://snellman.kootutteokset.fi/



Kuva 1. Venäjän sotavoimissa palvelleiden kenraalien ja amiraalien (vasemmalla) ja papiston (oikealla) elämänkaarien prosopografinen vertailu Biografiasampo.fi-palvelussa.

Haasteena ajallisen muutoksen hallinta

Keskeinen ontologinen haaste historiallisten paikkojen esittämisessä on, että ajan kuluessa paikkojen ominaisuudet muuttuvat. Jos haetaan vaikkapa Suomessa julkaistuja kirjoja, pitäisi Viipurissa painetut kirjat tulla hakutulokseen vain sotia edeltävältä ajalta.

Jo Suomen kuntien lähihistorian ajallinen paikkaontologia on varsin monimutkainen johtuen muutoksista kuntien laajentuessa, yhdistyessä, jakautuessa ja valtiorajojen muuttuessa.

Yhden hetkellisen ontologisen kuvauksen, kuten nykyisen Maanmittauslaitoksen Paikannimirekisterin (PNR) sijasta tarvitaan ontologioiden aikasarja, jossa on uusi paikkahierarkia periaatteessa jokaisen muutoskohdan jälkeen.

Aalto-yliopiston and Geologian tutkimuskeskuksen (GTK) kehittämä Suomen ajallinen paikkaontologia SAPO[14] kuntamuutoksista on yksi tapa lähestyä ongelmaa. Siinä luotiin uusi URI ja identiteetti jokaiselle ajalliselle erilaiselle kunnalle Suomessa 1800-luvulta lähtien aina, kun kunnan alue muuttui. Lisäksi luotiin oma tunniste koko kunnan käsitteelle eri aikoina. Käytettävissä ei ollut juurikaan tietoa kuntien polygoneista eri aikoina vaan ainoastaan kirjoissa dokumentoitua tietoa pinta-aloista ja muutoksista, kuten kuntaliitoksista. Tiedon perusteella muodostetusta ontologiasta voitiin kuitenkin päätellä alueellisia kattavuuksia tiedonhakua varten, kuten että vanhan Viipurin (ennen vuotta 1906) alueesta noin 12 prosenttia kuuluu edelleen Suomeen, nykyiseen Lappeenrantaan.

[14] https://seco.cs.aalto.fi/ontologies/sapo/

Suomalaista historiallisten paikkojen ontologiaa ja siihen liittyvää historiallisten karttojen palvelua ollaan tutkimassa ja kehittämässä Aalto-yliopiston ja Helsingin yliopiston Digitaalisten ihmistieteiden keskuksen HELDIG[15] yhteistyönä osana Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)[16] -kokonaisuutta.

Varteenotettava vaihtoehto ajallisen dimension tuomiseksi paikkaontologiaan on liittää sen ominaisuuksiin lisätietoa siinä, milloin kyseinen ominaisuus on ollut voimassa, esimerkiksi milloin Pietari oli Venäjän pääkaupunki. Näin on menetelty mm. semanttisesti rikkaimmassa kansainvälisessä historiallisten paikkojen rekisterissä, Getty-säätiön Thesaurus of Geographical Names (TGN) -ontologiassa[17], joka nykyisin julkaistaan avoimena linkitettynä datana verkossa. Suomessa kehitettävissä ontologioissa on hyvä huomioida kansainvälinen yhteentoimivuus paitsi W3C:n yleisiin semanttisen webin standardeihin myös TGN:n kaltaisiin sovellusalakohtaisiin de facto -standardeihin.

Eero Hyvönen on Helsingin yliopiston Digitaalisten ihmistieteiden keskuksen HELDIG johtaja ja Aalto-yliopiston tietotekniikan laitoksen professori. Hän on julkaissut yli 400 artikkelia ja kirjaa tekoälyyn ja semanttiseen webiin liittyen, uusimpana suomeksi ”Semanttinen web: avoimen linkitetyn datan käsikirja” (Gaudeamus, 2018).

Kotisivu ja yhteystiedot: https://seco.cs.aalto.fi/u/eahyvone/

[15] http://heldig.fi
[16] https://seco.cs.aalto.fi/projects/lodi4dh/
[17] http://www.getty.edu/research/tools/vocabularies/tgn/